240F Econometrics & Machine Learning Lecture 11-12 5/2/16

To date: $Y = f(X) + u$

Find $\hat{f}(X)$

Simple prediction. Done

Sometimes useful in econometrics.

Now: Consider workhorse models for causal inference in econometrics

(1) IV when instruments are known and valid but there are too many.

If we use them all have too many instruments problems

(2) Selection on observables model.

Consistent estimates once include controls but there are too many controls.

Solution: Select a subset of variables but

(1) Avoid overfitting / overtraining

(2) Include enough but not too many.

Survey article Belloni et al JEP Spring 2014 pp 29-50

240F Instrumental Variables & Lasso

- ~~Key~~ articles

  Full details: Belloni et al. Econometrica 2012 pp 2369-2429

  Survey: Belloni et al JEP Spring 2014 pp 39-41

- Consider

  model: $y_i = d_i'\alpha + \varepsilon_i$

  ↑ At least one component of d_i is endogenous

  Instruments: x_i satisfy $E[\varepsilon_i | x_i] = 0$

  The optimal instrument with i.i.d errors is

  $$D(x_i) = E[d_i | x_i]$$

  and leads to IV estimator with variance matrix $\sigma^2 \left\{ \frac{1}{n} \sum_{i=1}^{n} E[D(x_i) D(x_i)'] \right\}^{-1}$.

- Problem: Clearly best to use instruments that approximate $D(x_i)$

  eg. regress d_i on series expansion of x_i

  But this may lead to too many instruments so asymptotic theory poor.

  Solution: ~~Arbitrarily~~ pick just a few functions of x_i as instruments. But how? Lasso

  Use Lasso to pick instruments. Then IV as usual.

IV & LASSO

- n = sample size

  p = # instruments

  p large relative to n (even $p > n$) due to

  (1) many potential x's as instruments (many instrument)

  (2) a few x's but consider flexible model with powers and interactions. (many series instruments)

- ~~Use the lasso. For each endogenous regressor regress do lasso on the x's ~~ to select the subset of~~

- Let the p instruments be

  $f_i \equiv f(i_1, \ldots, f_{ip})' = (f_1(x_i), \ldots, f_p(x_i))'$

- Approximate sparsity

  Assume ~~approximate~~ that only s of these instruments are needed to provide a good approximation to $D_\ell(x_i) = E[d_{\ell i} | x_i]$ for the ℓ^{th} endog. variable

  i.e. $D_\ell(x_i) = f_i' \beta_{\ell 0} + a_\ell(x_i)$ for $\ell = 1, \ldots, k_e$

  $$\max_{1 \le \ell \le k_e} \sum_{j=1}^{p} 1[\beta_{\ell 0 j} \neq 0] \le s = o(n)$$

  where approximation error $a_\ell(x_i)$ is at most $O_p(\sqrt{s/n})$

lasso

- $d_{il} = D_l(x_i) + v_{il}$ $\quad E[v_{il}|x_i] = 0$ $\quad l = 1, \ldots, k_e$

  $\hat{D}_{il} = D_l(x_i) = f_i' \hat{\beta}_l \quad l = 1, \ldots, k_e$

- lasso $\hat{\beta}_{lL} = \underset{\beta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (d_{il} - f_i'\beta)^2 + \frac{\lambda}{n} \| \hat{\gamma}_l(\beta) \|_1$

  where λ is the penalty level

  $\hat{\gamma}_l = \operatorname{diag}(\hat{\gamma}_{l1}, \ldots, \hat{\gamma}_{lp})$ penalty loadings

  ideally $\hat{\gamma}_{lj} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} f_{ij}^2 v_{il}^2} \quad j = 1, \ldots, p$

- Since v_{il} is not observed start with conservative penalty loadings and then plug in the resulting estimated residuals, iterate.

  $\lambda = c\, 2\sqrt{n}\, \Phi^{-1}\left(1 - \frac{\gamma}{2kp}\right)$

  $\gamma = \frac{0.1}{\ln(\max(p, n))}$

  $c = 1.1$

- Once lasso picks the $\hat{\beta}_l \neq 0$

  post lasso does OLS with just these regressors.

  (OLS avoids shrinkage bias. Lasso avoids overfitting).

IV & LASSO

IV Form $\hat{D}_{il} = f_l(x_i)'\hat{\beta}_l \quad l = 1, \ldots, k_E$

$\hat{\beta}_l$ called be either Lasso $\hat{\beta}_l$ or post lasso $\hat{\beta}_l$

AND $\hat{D}_i = (\hat{D}_{i1}(x_i), \ldots, \hat{D}_{ik_E}(x_i), w_i')'$

↑ exogenous regressors

(Aside what if $\hat{D}_{il}(x_i)$ involves w_i ?)

IV $\hat{\alpha} = \left(\frac{1}{n}\sum_{i=1}^{n} \hat{D}_i \hat{d}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \hat{D}_i y_i\right)$

Can use usual asymptotic theory.

With heteroscedastic errors use vce (robust)

$$\hat{V}(\hat{\alpha}) = \left(\sum_{i=1}^{n} \hat{D}_i \hat{d}_i'\right)^{-1}\left(\sum_{i=1}^{n} \hat{\varepsilon}_i^2 \hat{D}_i \hat{D}_i'\right)^{-1}\left(\sum_{i=1}^{n} \hat{D}_i \hat{D}_i'\right)$$

① model selection mistakes are small enough to not matter

② lasso avoids overfitting.

Lasso & IV

Application

$$y_{ct} = \alpha_c + \alpha_t + \gamma_{ct} + \beta \text{ Takings law} + w_{ct}'\delta + \varepsilon_{ct}$$

y_{ct}: $\sim$ time t, judicial circuit c

Instruments arise due to random assignment of judges

# of takings appellate decisions that rule that a taking was unlawful

$n = 183$

$p = 167$

I think they partial out $\alpha_c, \alpha_t, \gamma_{ct}, w_{ct}'\delta$

So just $\hat{y}_{ct} = \alpha + \beta \widetilde{\text{Takings law}} + \text{error}$.

Focus on y_{ct} = Case-Shiller home price index within circuit c at time t

Extra paper finds $S = 2$ instruments

JEP " " $S = 1$

	OLS	2SLS	Lasso
$\hat{\beta}$	.0152	.0604	.0631
$se(\hat{\beta})$	.0132	.0296	.0249

lassoShooting y z*

Chen & Yeh (2010)

First stage F 89.6

First stage F 67.7

Control Function & Lasso

Full details Belloni et al. REStud 2014 608-650

Survey Belloni et al JEP spring 2014 41-45

- Consider partially linear model

$$y_i = d_i\alpha_0 + g(z_i) + \xi_i \qquad E[\xi_i | z_i, d_i] = 0$$

If z_i was low dimension can eg. use Robinson's 1988 differencing estimator.

- Instead there are many z_i:

In fact p potential regressors $x_i = f(z_i)$

- Assume enough sparsity so that only $s < n$ of x_i matter, approximating

$g(z_i)$ by linear combination of these

and $E[d_i | z_i] = m(z_i)$.

Control Function & Lasso

- $y_i = d_i \alpha_0 + g(z_i) + \xi_i \quad E[\xi_i | z_i, d_i] = 0$

  $d_i = m(z_i) + v_i \quad E[v_i | z_i] = 0$

Approximate by linear combinations of $x_i = p(z_i)$

$$y_i = d_i \alpha + \underbrace{x_i' \beta_g + r_{gi}}_{g(z_i)} + \xi_i$$

$$d_i = \underbrace{x_i' \beta_m + r_{mi}}_{m(z_i)} + v_i$$

$x_i = p(z_i)$ has dimension $p = p_n$ with $\log p = o(n^{1/3})$

/

B-splines, dummies, polynomials, interactions, ~

Sparsity

① $\|\beta_m\|_0 \le s$ or $\|\beta_g\|_0 \le s$ for $s << n$

↑ # non-zero β

② Approximation errors are small relative to estimation error

$$\sqrt{\sum_i E[r_{gi}^2]} \le c \times \sqrt{s/n} \quad \text{for some constant } c > 0$$

$$\sqrt{\sum_i E[r_{mi}^2]} \le c \times \sqrt{s/n}$$

240F Control Function & Lasso

Double Selection

- RF ① $y_i = x_i'\bar{\beta}_0 + \bar{r}_i + \bar{\xi}_i$ $\quad \bar{\beta}_0 = \alpha_0\beta_m + \beta_{g0}$

  ② $d_i = x_i'\beta_m + r_{mi} + v_i$ $\quad \bar{r}_i = \alpha r_{mi} + r_{gi}$

  $\bar{\xi}_i = \alpha v_i + \xi_i$

  ↑ sub out d_i in first eqn.

- Apply lasso to each equation separately!

  Similar λ penalty to Ecta paper.

  Then choose x_i to be union of x's from ① & ② (Double selection).

- Then do OLS of y_i on d_i and these x_i's.

  Can apply regular asymptotics

- Why double selection?

  Will sometimes drop a control variable in the original equation when $\beta_g \simeq 0$. But it still could be really important if $\beta_m \neq 0$ and $\alpha \neq 0$.

240F

Control Function & Lasso

- Paper section 5 discusses heterogeneous effects.
  Also see Max Farrell 2015 J. Econometrics 189 1-23

- Application

Donohue & Levitt Abortion and Crime

$$y_{it} = \alpha a_{it} + w_{it}'\beta + \delta_i + \gamma_t + \varepsilon_{it}$$

y_{it}: Crime violent property murder

a_{it}: Abortion rate

w_{it}: 264 controls plus time effects that are partialled out

$n = 48$
$T = 12$
$nT = 576$

or

$n = 50$
$T = 12$
$nT = 600$

Propensity Score Matching

Doubly Robust Estimator of ATE

- Max Farrell 2015 J. Econometrics 189 1-23
- Multivalued treatment $D \in \{0, 1, \ldots, T\}$

Generalized treatment score $p_t(x) = Pr[D = t \mid X = x]$

Conditional outcome mean, $E[Y \mid D = t, X = x] =$ function $\mu_t(x) =$

iid sample $\{y_i, d_i, x_i\}$ $i = 1, \ldots, n$

Assume selection on observables

$$\hat{\mu}_t = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{1(d_i = t)(y_i - \hat{\mu}_t(x_i))}{\hat{p}_t(x_i)} + \hat{\mu}_t(x_i) \right\}$$

1. Fit p_t & μ_t using group lasso applied to multinomial logit & LS
2. Refit p_t, μ_t using unpenalized models possibly augmented with controls from theory / prior work

Apply to Dehejia Wahba (1999) like Robins, Rotnitzky (1995)

Grouped lasso ESL p. 90 with L groups (γ is $L \times 1$)

$$\min_{\beta} \left(\left\| y - \beta_0 \underset{\sim}{e} - \sum_{l=1}^{L} X_l \beta_l \right\|_2^2 + \lambda \sum_{l=1}^{L} \sqrt{p_l} \, \|\beta_l\|_2 \right)$$

$\| \ \|_2$ denotes Euclidean norm without squaring

$$\|a\|_2 = \sqrt{a'a}$$

2 of ~~4~~ 2

BIG DATA

- Hal Varian Big Data: New Tricks for Econometrics
  JEP Spring 2014 3-28

- Tools
  - Hadoop File System — Large Files
  - Google File System → Cassandra — Table of data that can stretch over many computers
  - Big Table → Hadoop (Google MapReduce) — System to access & manipulate data in large data structure
  - mapReduce → Pig — Language to create mapreduce jobs
  - Sawzall → None — Language for parallel data process
  - Go
  - Dremel (Big Query), HM (Drill) — SQL queries

- Penalized regression
  Train/test/validate
  K-fold CV
  Classification with regression trees

Further work
Imbens & Athey